Об одном подходе к реализации алгоритмов Нидлмана – Вунша и Джаро – Винклера и их применении в корреляционном анализе сходства митохондриальных ДНК обезьян. Часть II. Вычислительные эксперименты

Б. Ф. Мельников, Ли Цзямянь, Му Цзинъюань

***Аннотация*—В исследованиях молекулярной биологии и геномики крайне важно понимать генетические различия между разными видами. Сравнение сходства последовательностей ДНК может предоставить ценную информацию о родственных связях между видами.**

В настоящей статье для сравнения митохондриальных ДНК обезьян использовались два алгоритма – Нидлмана – Вунша и Джаро – Винклера; кроме того, в последующих частях статьи будут приведены подобные сравнения и для других млекопитающих.

Ранее при проведении подобных исследований у настоящей статьи авторов возникла гипотеза о том, что при применении этих двух алгоритмов для анализа сходства одних и тех же пар геномных последовательностей получаются весьма непохожие результаты. Одним из предметов настоящей статьи и является описание подхода к тому, как именно мы предлагаем давать численные ответы на подобные вопросы. Такие ответы мы предполагаем давать с помощью применения ранговой корреляции, о которой будет сказано в следующих частях статьи.

Из результатов настоящей статьи следует необходимость продолжения подробных исследований цепочек ДНК, в частности, на предмет анализа их сходства; то есть подобные задачи остаются и ещё на долгое время останутся весьма актуальными.

Основное содержание второй части статьи – описание выполненных вычислительных экспериментов.

***Ключевые слова*—последовательности ДНК, алгоритм Нидлмана – Вунша, алгоритм Джаро – Винклера, матрица расстояний, корреляционный анализ.**

В конце первой части настоящей статьи [1] был приведён такой текст (здесь даётся с сокращением): «основной предмет последующих частей настоящей статьи – описание инструментария для подсчёта ранговой корреляции между значениями badness для последовательностей треугольников, полученных на основе матриц расстояний между цепочками мт ДНК …» Однако сам этот инструментарий мы решили изложить в виде отдельной публикации, мало связанной с исследованием ДНК-цепочек: [2]; поэтому реальный объём второй части небольшой.

Статья получена 5 декабря 2024 г.

Борис Феликсович Мельников, Университет МГУ – ППИ в Шэньчжэне (bormel@smbu.edu.cn).

Ли Цзямянь, Университет МГУ – ППИ в Шэньчжэне (lijiamian0804@live.com).

Му Цзинъюань, Университет МГУ – ППИ в Шэньчжэне (xirousang@gmail.com).

Мы продолжаем нумерацию разделов, рисунков и таблиц; нумерация библиографических ссылок новая.

Приведём содержание части II по разделам. В разделе VI представлены некоторые результаты сравнения алгоритмов расчёта расстояний между ДНК на основе треугольной нормы. Полное (более подробное, чем использованное) название раздела VII могло бы быть таким: «Обобщение результатов сравнения алгоритмов для расчёта расстояний между ДНК на основе ранговой корреляции» – такое название полностью описывает содержание этого раздела. Раздел VIII – заключение; в нём мы кратко описываем возможные направления будущих работ, связанных с рассмотренной в статье тематикой.

VI. Результаты сравнения алгоритмов вычисления расстояний между ДНК

В этом разделе мы представляем некоторые результаты сравнения алгоритмов расчета расстояний между ДНК на основе треугольной нормы [1]. Мы используем только одну из норм, рассмотренных выше. Однако, было бы точнее сказать, что мы представляем *подход к получению таких результатов*; этот подход основан на применении соответствующих показателей, описанных в этом разделе.

Из рассмотренных выше треугольных норм (badness) мы будем использовать только одну – номер (0) из таблицы II первой части [1]; она кажется несколько более адекватной. Именно для неё мы приводим некоторые результаты сравнения алгоритмов расчёта расстояний между ДНК.

Ещё раз отметим, что пример, приведённый в части I для человека, шимпанзе и бонобо, *может быть обобщён на любые три вида*; поэтому и можно рассматривать треугольную норму, определённую там. В то же время мы можем добавить следующее размышление. Конечно, мутации в геномах накапливаются более или менее пропорционально количеству прошедших поколений; однако очевидно, что у людей их меньше, чем за тот же период у шимпанзе. Но условно, поскольку геномы шимпанзе и

[1] Название «треугольная метрика» вряд ли удачное: в предыдущих публикациях уже говорилось о том, что при подсчётах изредка нарушается неравенство треугольника. Впрочем, во-первых, в идеальном случае оно не нарушается, и, во-вторых, в предыдущих публикациях слово «метрика» мы обычно писали в кавычках.

бонобо отделились от генома человека одновременно, то число их мутаций должно быть ближе друг к другу, чем от них обоих к человеку. Это дополнение также косвенно объясняет возможность использования треугольной нормы.

По-видимому, в наиболее ярком виде обобщение этого примера можно продемонстрировать именно для всех человекообразных обезьян (включая гиббона – мы для примера будем рассматривать только один вид, как и считалось до 2005 г.[2], – хотя сейчас в наиболее признанной классификации число видов в этом роду равно 17); см. следующий рисунок 4:

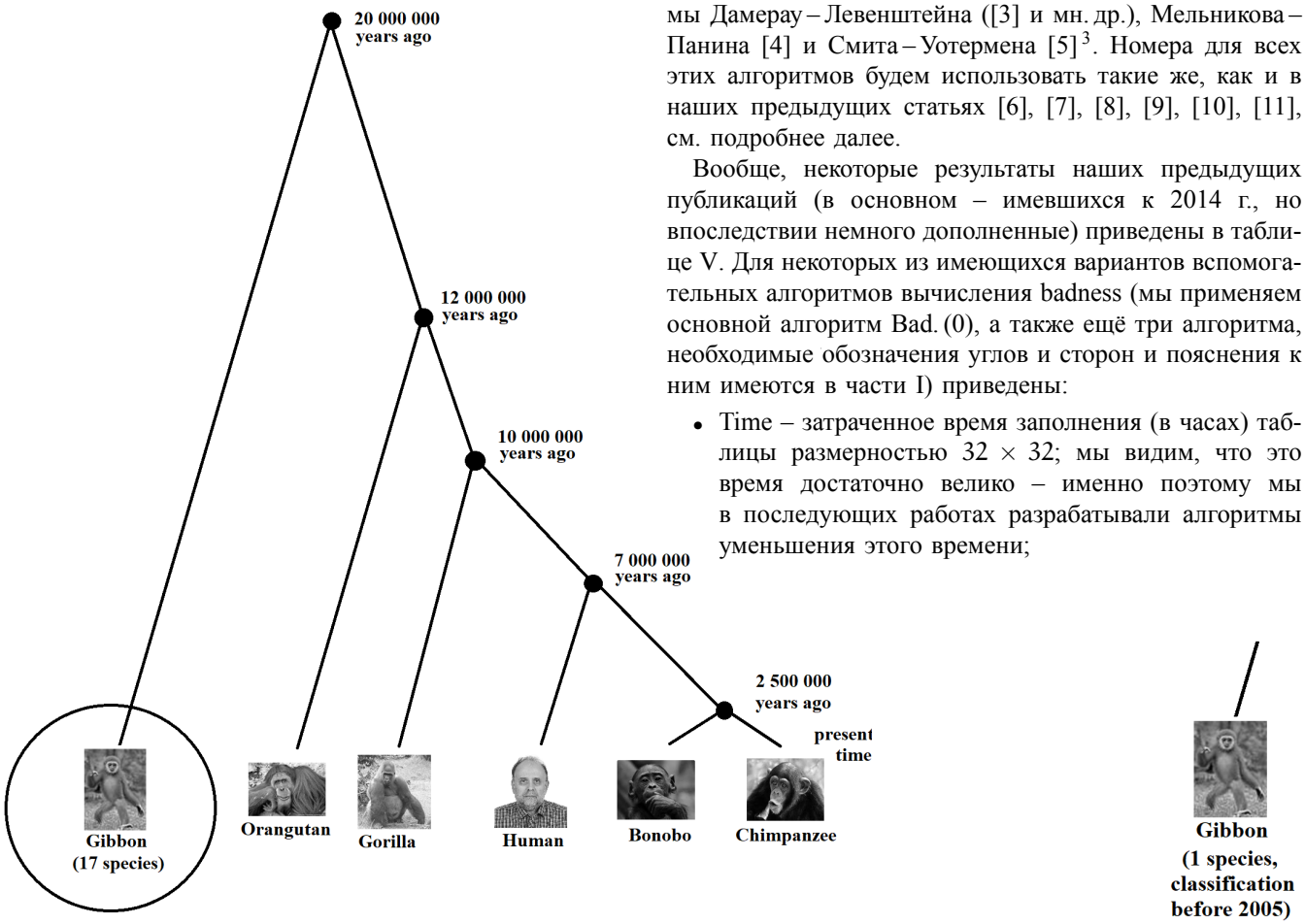

Рис. 4. Дерево классификации высших обезьян, включая человека (ви́дение классификации биологами: современное и недавнее)

Приведённый рисунок 4 интересен в первую очередь тем, что дерево разделения видов (двоичное!) фактически превратилось на нём в список (однонаправленный) – что, по-видимому, бывает достаточно редко. Однако, с точки зрения авторов статьи, это в ещё большей степени иллюстрирует как обоснованность применения «треугольной метрики» (вычисляемые нами значения badness), так и возможность расширения применяемых эвристик.

Вернёмся непосредственно к алгоритмам расчёта расстояний между ДНК-цепочками; напомним, что в название статьи вынесены алгоритмы Нидлмана – Вунша и Джаро – Винклера (подробное описание нашей их реализации приведено в части I) – и «добавим» к ним алгоритмы Дамерау – Левенштейна ([3] и мн. др.), Мельникова – Панина [4] и Смита – Уотермена [5][3]. Номера для всех этих алгоритмов будем использовать такие же, как и в наших предыдущих статьях [6], [7], [8], [9], [10], [11], см. подробнее далее.

Вообще, некоторые результаты наших предыдущих публикаций (в основном – имевшихся к 2014 г., но впоследствии немного дополненные) приведены в таблице V. Для некоторых из имеющихся вариантов вспомогательных алгоритмов вычисления badness (мы применяем основной алгоритм Bad. (0), а также ещё три алгоритма, необходимые обозначения углов и сторон и пояснения к ним имеются в части I) приведены:

- Time – затраченное время заполнения (в часах) таблицы размерностью 32×32; мы видим, что это время достаточно велико – именно поэтому мы в последующих работах разрабатывали алгоритмы уменьшения этого времени;
- Vio – число нарушений в этой таблице неравенства треугольника (некоторые подробности см. ниже);
- Bad. (0), …, Bad. (3) – полученные значения badness для 4 вспомогательных алгоритмов её вычисления.

Таблица V
Средние значения badness для разных методов

No.	Time (h.)	Vio.	Bad. (0) $(\alpha-\beta)/\gamma$	Bad. (1) $(\alpha-\beta)/\pi$	Bad. (2) $(\alpha-\beta)/\alpha$	Bad. (3) $(a-b)/a$
1, D–L	27	0	**0.155**	**0.0522**	**0.121**	0.0527
2, N–W	2.1	0	**0.101**	**0.0314**	**0.0692**	**0.0290**
3, J–W	2.3	0	1.331	0.501	0.600	0.154
4, M–P	28	12	**0.155**	0.0527	0.122	**0.0482**
5, S–W	28	14	0.200	0.0732	0.150	0.0608

[2] Более того, согласно одной из принятой до 2005 г. классификаций, этот единственный вид *включался в подсемейство человекообразных обезьян* – которое включало, конечно, и самого́ человека. Приведённый рисунок в большой степени отражает такую классификацию. Не имеет никакого значения, что *именно эту* классификацию уже «отменили»: для каких-то других видов что-то аналогичное принято и сейчас.

[3] Важно отметить, что в Интернете имеется очень немного *приемлемых для практического применения* реализаций алгоритмов вычисления расстояний между ДНК-цепочками. Мы рассматриваем 5 алгоритмов – но при этом один принадлежит одному из авторов статьи (с соавтором), ещё два алгоритма нами были переписаны заново (см. часть I), так что реально мы применяли только два «чужих» алгоритма.

Добавим конкретные примечания про нарушения неравенства треугольника – мы считаем, что это далеко не самая важная характеристика для выбранного алгоритма вычисления badness; это можно объяснить следующим:

- во-первых, треугольников в матрице расстояний размерности 32 всего получается
$$\frac{32 \cdot 31 \cdot 30}{2 \cdot 3} = 4\,960,$$
так что даже в случае «самого неудачного» алгоритма «ненастоящий» треугольник появляется реже чем в 1 случае из 350;
- во-вторых, в формуле для окончательного значения badness мы уже учли такие возможные «ненастоящие» треугольники – и в самом плохом случае такое значение badness полагаем равным 2; подробности см. в процитированных выше статьях.

Перейдём к описанию результатов, приведённых в таблице. В каждом её столбце выделены по 2 лучших значения – что *при условном ранжировании* алгоритмов ставит на первое место алгоритм Нидлмана – Вунша, и примерно одинаково на второе–третье места алгоритмы Мельникова – Панина и Дамерау – Левенштейна [4]. Также эта же таблица может показать, что мы *уже заранее* (т. е. только приступая к основному исследованию этой статьи) были плохого мнения об алгоритме Джаро – Винклера – что и подтвердилось результатами статьи (косвенно).

В заключение раздела приведём ещё один косвенный аргумент – тоже о том, что весь наш подход, связанный с исследованием матриц расстояний на основе значений badness всех треугольников, является адекватным рассматриваемым задачам. Во многих наших предыдущих публикациях мы для среднего значения badness некоторой матрицы подбирали треугольник:

- со сторонами, выраженными натуральными числами, образующими арифметическую прогрессию с разностью 1;
- причём такой из них, чтобы его badness минимально отличалась бы от среднего значения badness рассматриваемой матрицы.

В нашей ситуации к наилучшему значению Bad. (0) таблицы, примерно равному 0.101, наиболее близок треугольник со сторонами 19, 18 и 17 [5]; по нашему мнению, этот треугольник визуально почти не отличается от равностороннего.

VII. ОБОБЩЕНИЕ РЕЗУЛЬТАТОВ СРАВНЕНИЯ НА ОСНОВЕ РАНГОВОЙ КОРРЕЛЯЦИИ

Полное (более подробное, чем использованное) название этого раздела VII могло бы быть таким: «Обобщение результатов сравнения алгоритмов для расчёта расстояний между ДНК на основе ранговой корреляции» – такое название полностью описывает содержание раздела.

Как уже было сказано выше, сам инструментарий, связанный с алгоритмами ранговой корреляции, а также со сравнением алгоритмов расчёта расстояний между строками ДНК на её основе, мы решили изложить в виде отдельной публикации, практически не связанной с исследованием ДНК-цепочек: [2].

Приведём по этому поводу только два пояснительных рисунка.

На первом из них (рис. 5) мы показали соответствующие треугольники в двух разных матрицах: считаем, что матрица A выполнена для какого-то одного алгоритма, а матрица B – для какого-то другого.

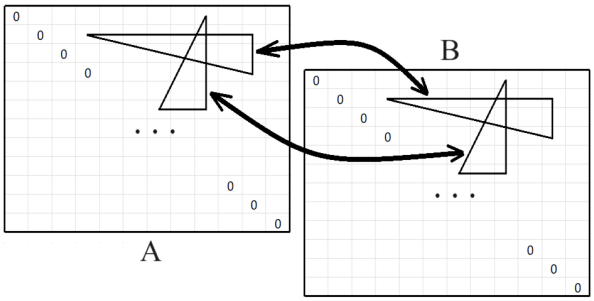

Рис. 5. Соответствующие треугольники в двух разных матрицах

Мы должны сравнить значения badness для этих треугольников – и в идеальном случае всегда (для всех пар) должны получаться одинаковые результаты сравнения (то есть либо оба знака суть $<$, либо оба знака $>$, либо оба знака $=$).

	N-W	D-L
1	0	0
2	0	0
3	0	0
	• • •	
13	0	0
14	0.017	0.0095
15	0.017	0.0096
	• • •	
117	0.029	0.0221
118	0.030	0.0219
	• • •	
5981	0.163	0.173
5982	0.164	0.173
5983	0.165	0.175
5984	0.181	0.177

Рис. 6. Процесс работы алгоритма вычисления ранговой корреляции

На втором (рис. 6) мы привели процесс работы алгоритма вычисления ранговой корреляции по методу Кендалла для реальных значений badness треугольников. На основе проделанных вычислений мы выяснили, что наиболее удачный рисунок получается для примера, сравнивающего алгоритмы Нидлмана – Вунша и Дамерау – Левенштейна (вторым из этих алгоритмов мы на рисунке заменили алгоритм Джаро – Винклера, являющийся одним из предметов настоящей статьи), применительно к 34 видам обезьян (а не 32, как в остальных примерах статьи).

[4] Всё-таки мы считаем, что наш алгоритм *совсем немного* лучше – и это видно, в частности, в приведённой таблице. При этом оба алгоритма достаточно близки по своей идеологии – именно поэтому большинство значений в соответствующих столбцах почти одинаково.

[5] Его badness примерно равна 0.1097, или в точности 0.101 – если использовать такое число значащих цифр, которое употреблено в приведённой таблице.

Таблица VI
Основные результаты вычислений

Option	corr-0, usual	corr-1, Spearman	corr-2, Kendall+	corr-3, Kendall++	corr-4, our
without	0.0817	0.136	0.0742	0.0909	–
with	0.0817	0.136	0.139	0.0909	0.106

Конкретно, на рисунке показан первый по номерам случай «энтропии» – т. е. нарушения порядка значений badness для последовательностей их значений, где первая из последовательностей упорядочена в порядке неубывания (а при совпадении значений – мы упорядочиваем пары по порядку неубывания для второй последовательности). Итоговое значение коэффициента Кендалла, как следует из описания соответствующего алгоритма, линейно зависит от числа таких нарушений.

Итак, приведём итоговые результаты вычислений, а также сформулируем выводы, получаемые на основе этих результатов.

В целом, как следует из предыдущего материала, мы можем работать с таблицами III и IV (они были приведены в первой части статьи), а также с любыми другими таблицами, построенными по тому же принципу, просто как с *таблицами целых чисел*: интересующие нас значения badness будут одинаковыми.

Все основные результаты вычислений кратко даны в таблице VI. Заголовки столбцов понятны на основе предыдущего материала; во второй строке (“with”) мы использовали нормализацию, а в первой строке (“without”) её не использовали. При этом, как обычно, нормализацией мы называем линейное отображение всех полученных данных в какой-либо сегмент, например $[0, 1]$.

Итак, в статье мы показали очень небольшую зависимость между двумя хорошо известными алгоритмами определения расстояний между геномами – а именно, алгоритмами Джаро – Винклера и Нидлмана – Вунша. В частности, у авторов имеется предположение (пока ещё нами полностью экспериментально не подтверждённое), что алгоритм Нидлмана – Вунша значительно более адекватен, чем алгоритм Джаро – Винклера – то есть он даёт правильные (или очень близкие к ним) расстояния между геномами.

VIII. ЗАКЛЮЧЕНИЕ

Как уже отмечалось. в последующей работе стоит рассмотреть улучшение алгоритма Джаро – Винклера – для динамической адаптации диапазона префиксов к длине конкретной последовательности гена, а также в некоторых ситуациях сделать возможным использование отрицательных значений весов – чтобы тем самым учитывать снижение или повышение веса префикса в общей оценке сходства геномов.

БЛАГОДАРНОСТИ

Настоящая работа была частично поддержана грантом научной программы китайских университетов “Higher Education Stability Support Program” (раздел “Shenzhen 2022 – Science, Technology and Innovation Commission of Shenzhen Municipality”) – 深圳市 2022 年高等院校稳定支持计划资助项目.

Список литературы

[1] Ли Цзямянь, Му Цзинъюань, Мельников Б.Ф. Об одном подходе к реализации алгоритмов Нидлмана – Вунша и Джаро – Винклера и их применении в корреляционном анализе сходства митохондриальных ДНК обезьян. Часть I. Общее описание работы // International Journal of Open Information Technologies. 2024. Vol. 12, No. 9. P. 1–10.

[2] Мельников Б.Ф. Об одном подходе к вычислению ранговой корреляции. Часть I // International Journal of Open Information Technologies. 2024. Vol. 12, No. 11. P. 1–8.

[3] D-L Levenshtein V. Binary codes capable of correcting. Deletions, insertions, and reversals // Soviet Physics Doklady. 1966. Vol. 10. P. 707–710.

[4] Мельников Б.Ф., Панин А.Г. Параллельная реализация мультиэвристического подхода в задаче сравнения генетических последовательностей // Вектор науки Тольяттинского государственного университета. 2010. № 4 (22). С. 83–86.

[5] Munekawa Y., Ino F., Hagihara K. Design and implementation of the Smith-Waterman algorithm on the CUDA-compatible GPU. // 8th IEEE International Conference on BioInformatics and BioEngineering, BIBE-2008. DOI: 10.1109/BIBE.2008.4696721.

[6] Мельников Б.Ф., Тренина М.А., Кочергин А.С. Подход к улучшению алгоритмов расчета расстояний между цепочками ДНК (на примере алгоритма Нидлмана – Вунша) // Известия высших учебных заведений. Поволжский регион. Физико-математические науки. 2018. № 1 (45). С. 46-59.

[7] Мельников Б.Ф., Тренина М.А. Об одной задаче восстановления матриц расстояний между цепочками ДНК // International Journal of Open Information Technologies. 2018. Vol. 6, No. 6. P. 1–13.

[8] Абрамян М.Э., Мельников Б.Ф., Тренина М.А. Реализация метода ветвей и границ для задачи восстановления матрицы расстояний между последовательностями ДНК // Современные информационные технологии и ИТ-образование. 2019. Т. 15, № 1. С. 81–91.

[9] Melnikov B., Chaikovskii D. Some general heuristics in the traveling salesman problem and the problem of reconstructing the DNA chain distance matrix // ACM International Conference Proceeding Series. 2023. P. 361–368.

[10] Abramyan M., Melnikov B., Zhang Y. Some more on restoring distance matrices between DNA chains: reliability coefficients // Cybernetics and Physics. 2023. Vol. 12, No. 4. P. 237–251.

[11] Melnikov B., Chaikovskii D. On the Application of Heuristics of the TSP for the Task of Restoring the DNA Matrix // Frontiers in Artificial Intelligence and Applications. 2024. Vol. 385. P. 36–44.

Борис Феликсович МЕЛЬНИКОВ,
профессор Университета МГУ – ППИ в Шэньчжэне
(http://szmsubit.ru/),
email$_1$: bormel@smbu.edu.cn,
email$_2$: bf-melnikov@yandex.ru,
mathnet.ru: personid=27967,
elibrary.ru: authorid=15715,
scopus.com: authorId=55954040300,
ORCID: orcidID=0000-0002-6765-6800.

ЛИ Цзямянь,
аспирант Университета МГУ – ППИ в Шэньчжэне
(http://szmsubit.ru/),
email: lijiamian0804@live.com.

МУ Цзинъюань,
аспирант Университета МГУ – ППИ в Шэньчжэне
(http://szmsubit.ru/),
email: xirousang@gmail.com.

On an approach to the implementation of the Needleman – Wunsch and Jaro – Winkler algorithms and their application in the correlation analysis of the similarity of mitochondrial DNA of monkeys. Part II

Boris Melnikov, Li Jiamian, Mu Jingyuan

***Abstract*—In molecular biology and genomics research, it is very important to understand the genetic differences between different species. Comparing the similarity of DNA sequences can provide valuable information about the relationships between species.**

In this paper, two algorithms were used to compare the mitochondrial DNA of monkeys, i.e., Needleman – Wunsch and Jaro – Winkler algorithms. In addition, in the following parts of the paper, similar comparisons will be made for other mammals.

Earlier, when conducting such studies, the authors of this paper had a following hypothesis. When using these two algorithms to analyze the similarity of the same pairs of genomic sequences, very different results are obtained. One of the subjects of this paper is a description of the approach to how exactly we propose to give numerical answers to such questions. We propose to give such answers using the use of pair correlation, which will be discussed in the following parts of the paper.

From the results of this paper, it follows that it is necessary to continue detailed studies of DNA chains, in particular, to analyze their similarity. That is, such problems remain and will remain very relevant for a long time.

The main content of the second part of the paper is a description of the performed computational experiments.

***Keywords*—DNA sequences, Needleman–Wunsch algorithm, Jaro–Winkler algorithm, distance matrix, correlation analysis.**

References

[1] Li Jiamian, Mu Jingyuan, Melnikov B. On an approach to the implementation of the Needleman –Wunsch and Jaro –Winkler algorithms and their application in the correlation analysis of the similarity of mitochondrial DNA of monkeys. Part I. The general description of the work // International Journal of Open Information Technologies. 2024. Vol. 12, No. 9. P. 1–10 (in Russian).

[2] Melnikov B. On one approach to calculating rank correlation. Part I // International Journal of Open Information Technologies. 2024. Vol. 12, No. 11. P. 1–8 (in Russian).

[3] Levenshtein V. Binary codes capable of correcting. Deletions, insertions, and reversals // Soviet Physics Doklady. 1966. Vol. 10. P. 707–710.

[4] Melnikov B., Panin A. Parallel implementation of the multiheuristic approach in the task of comparing genetic sequences // Vector of science of Tolyatti State University. 2022. No. 4 (22). P. 83–86 (in Russian).

[5] Munekawa Y., Ino F., Hagihara K. Design and implementation of the Smith-Waterman algorithm on the CUDA-compatible GPU. // 8th IEEE International Conference on BioInformatics and BioEngineering, BIBE-2008. DOI: 10.1109/BIBE.2008.4696721.

[6] Melnikov B., Trenina M., Kochergin A. An approach to improving algorithms for calculating distances between DNA chains (using the Needleman – Wunsch algorithm as an example) // News of higher educational institutions. Volga region. Physical and mathematical sciences. 2018. No. 1 (45). P. 46-59 (in Russian).

[7] Melnikov B., Trenina M. On a problem of reconstructing distance matrices between DNA chains // International Journal of Open Information Technologies. 2018. Vol. 6, No. 6. P. 1–13 (in Russian).

[8] Abramyan M., Melnikov B., Trenina M. Implementation of the branch and boundary method for the task of reconstructing the matrix of distances between DNA sequences // Modern information technologies and IT education. 2019. Vol. 15, No 1. P. 81–91 (in Russian).

[9] Melnikov B., Chaikovskii D. Some general heuristics in the traveling salesman problem and the problem of reconstructing the DNA chain distance matrix // ACM International Conference Proceeding Series. 2023. P. 361–368.

[10] Abramyan M., Melnikov B., Zhang Y. Some more on restoring distance matrices between DNA chains: reliability coefficients // Cybernetics and Physics. 2023. Vol. 12, No. 4. P. 237–251.

[11] Melnikov B., Chaikovskii D. On the Application of Heuristics of the TSP for the Task of Restoring the DNA Matrix // Frontiers in Artificial Intelligence and Applications. 2024. Vol. 385. P. 36–44.

Boris MELNIKOV,
Professor of Shenzhen MSU – BIT University, China
(http://szmsubit.ru/),
email$_1$: bormel@smbu.edu.cn,
email$_2$: bf-melnikov@yandex.ru,
mathnet.ru: personid=27967,
elibrary.ru: authorid=15715,
scopus.com: authorId=55954040300,
ORCID: orcidID=0000-0002-6765-6800.

LI Jiamian,
Post-graduate student of Shenzhen MSU – BIT University, China (http://szmsubit.ru/),
email: lijiamian0804@live.com.

MU Jingyuan,
Post-graduate student of Shenzhen MSU – BIT University, China (http://szmsubit.ru/),
email: xirousang@gmail.com.